Weakly-Supervised Camouflaged Object Detection with Scribble Annotations

Ruozhen He\textsuperscript{*}, Qihua Dong\textsuperscript{*}, Jiaying Lin\textsuperscript{†}, Rynson W.H. Lau\textsuperscript{†}

Department of Computer Science, City University of Hong Kong
\{ruozhenhe2-c, qihuadong2-c, jiayinlin5-c\}@my.cityu.edu.hk, Rynson.Lau@cityu.edu.hk

Abstract

Existing camouflaged object detection (COD) methods rely heavily on large-scale datasets with pixel-wise annotations. However, due to the ambiguous boundary, annotating camouflaged objects pixel-wisely is very time-consuming and labor-intensive, taking \(\sim\)60mins to label one image. In this paper, we propose the first weakly-supervised COD method, using scribble annotations as supervision. To achieve this, we first relabel 4,040 images in existing camouflaged object datasets with scribbles, which takes \(\sim\)10s to label one image. As scribble annotations only describe the primary structure of objects without details, for the network to learn to localize the boundaries of camouflaged objects, we propose a novel consistency loss composed of two parts: a cross-view loss to attain reliable consistency over different images, and an inside-view loss to maintain consistency inside a single prediction map. Besides, we observe that humans use semantic information to segment regions near the boundaries of camouflaged objects. Hence, we further propose a feature-guided loss, which includes visual features directly extracted from images and semantically significant features captured by the model. Finally, we propose a novel network for COD via scribble learning on structural information and semantic relations. Our network has two novel modules; the local-context contrasted (LCC) module, which mimics visual inhibition to enhance image contrast/sharpness and expand the scribbles into potential camouflaged regions, and the logical semantic relation (LSR) module, which analyzes the semantic relation to determine the regions representing the camouflaged object. Experimental results show that our model outperforms relevant SOTA methods on three COD benchmarks with an average improvement of 11.0\% on MAE, 3.2\% on S-measure, 2.5\% on E-measure, and 4.4\% on weighted F-measure.

Introduction

Camouflaged object detection (COD) aims to detect visually inconspicuous objects in their surroundings, which includes natural objects with protective coloring, small sizes, occlusion, and artificial objects with information hiding purposes. Ambiguous boundaries between objects and backgrounds make it a more challenging task than other object detection

\textsuperscript{*}These authors contributed equally.
\textsuperscript{†}Corresponding authors: Jiaying Lin and Rynson W.H. Lau
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Figure 2 shows the percentage of annotated pixels in S-COD. Compared to pixel-wise annotation, the labeling process of S-COD is much easier. Compared with other labeling approaches (e.g., box and point annotation), it provides more pixel-level guidance, allowing semantic information to be exploited, and is comparably efficient in labeling.

Nevertheless, how to exploit scribble annotations for COD is still under exploration. Directly applying existing scribble-based salient object detection (SOD) methods are not appropriate here since camouflaged objects are not salient. Figure 3 shows that two state-of-the-art scribble-based SOD methods, SS (Zhang et al. 2020b) and SCWSSOD (Yu et al. 2021), fail in two common scenarios. The first row of Figure 3 shows an object with an ambiguous boundary in the generally consistent background. Due to the similar low-level features, both SS and SCWSSOD experience difficulties recognizing the boundaries. The second row requires detectors to identify semantic relations of objects (e.g., flower stems and petals), as more than one object looks like the “camouflaged” foreground. Here, both SS and SCWSSOD mistakenly include other objects as the foreground, due to poor semantic information learning.

In this paper, we present the first scribble-based COD learning framework to address the weakly-supervised COD problem with scribble annotations. We observe that humans would first identify possible foreground objects (Wald 1935) and then use semantic information to exactly segment them (Hubel and Wiesel 1962). To incorporate this process in our model, we propose a feature-guided loss, which considers not only visual affinity but also high-level semantic features, to guide the segmentation. The high-level features are learned in an end-to-end fashion during training and do not depend on other well-trained detectors. In addition, in our network design, we propose the local-context contrasted (LCC) module to mimic visual inhibition in strengthening contrast (Von Békésy 2017) in order to find potential camouflaged regions, and the logical semantic relation (LSR) module to determine the final camouflaged object regions. Further, we notice that current weakly-supervised methods tend to have inconsistent predictions in COD, possibly due to the “camouflage” characteristics. Hence, we design a consistency regularization, which is stronger and more reliable than previous weakly-supervised learning methods. Specifically, we introduce the reliability bias in the cross-view loss to improve the self-consistency mechanism. We also present the inside-view consistency loss to reduce the uncertainty of predictions. The regularization enhances the stability and quality of the prediction.

In conclusion, our main contributions are as follows:

- We propose the first weakly-supervised COD dataset with scribble annotation. Compared with pixel-wise annotation, it takes only ~10 seconds to annotate each image (360 times faster) and overcomes the limitation of assigning equal importance to every object pixel.
- We propose the first end-to-end weakly-supervised COD framework. It includes novel feature-guided loss functions and consistency loss. Imitating human perceptions, the loss functions guide the network to extract high-level features that help distinguish objects and impose stability on the predictions.
- We propose a novel network for scribble learning, which utilizes low-level contrasts to expand the scribbles to wider camouflaged regions and logical semantic information to finalize the objects.
- Experimental results show that our framework outperforms relevant state-of-the-art methods on three COD benchmarks with an average improvement of 11.0% on MAE, 3.2% on S-measure, 2.5% on E-measure, and 4.4% on weighted F-measure.

**Related Work**

**Camouflaged Object Detection**. COD focuses on undetectable natural and artificial objects. (e.g., objects with similar appearances to the surrounding) (Fan et al. 2020a) proposes a COD dataset with 10K camouflaged images, which takes an average of around 60 minutes to annotate each image. (Zhai et al. 2021) proposes a mutual graph learning method that splits the task into rough positioning and precise boundary locating. (Li et al. 2021) applies joint learning on SOD and COD tasks, taking advantage of both tasks to meet a balance of global/local information. (Mei et al. 2021) proposes a focus module to detect and remove false-positive and false-negative predictions. (Yang et al. 2021) proposes a transformer-based probabilistic representational model to...
learn context information to solve uncertainty-guided ambiguity. (Lin et al. 2022) proposes a frequency-aware COD method. (Youwei et al. 2022) proposes a multi-scale network that employs the zoom strategy to learn mixed-scale semantics for accurate segmentation.

However, these methods highly rely on per-pixel ground-truth with full supervision, which is time-consuming and labor-intensive. To overcome these limitations, we propose scribble annotations to construct COD datasets, the first weakly-supervised dataset for COD task to our knowledge.

**Methodology**

The training dataset is defined as $D = \{x_n, y_n\}_{n=1}^{N_{img}}$, where $x_n$ is the input, $y_n$ is the annotation map, and $N_{img}$ is the total number of training images. In our task, $y_n$ is in scribble-form, in which 1 is foreground, 2 is background, and 0 is unknown pixels.

**Overall Structure**

The training dataset is defined as $D = \{x_n, y_n\}_{n=1}^{N_{img}}$, where $x_n$ is the input, $y_n$ is the annotation map, and $N_{img}$ is the total number of training images. In our task, $y_n$ is in scribble-form, in which 1 is foreground, 2 is background, and 0 is unknown pixels.

**Methodology**

The training dataset is defined as $D = \{x_n, y_n\}_{n=1}^{N_{img}}$, where $x_n$ is the input, $y_n$ is the annotation map, and $N_{img}$ is the total number of training images. In our task, $y_n$ is in scribble-form, in which 1 is foreground, 2 is background, and 0 is unknown pixels.

**Logical Semantic Relation (LSR) Module**

Scribble annotation may only annotate a part of the background. When the background consists of many low-level contrasted parts (e.g., green leaves and brown branches, yellow petals, and green stems), we need logical semantic relation information to identify the real foreground and background. Hence, we propose the LSR module to extract semantic features from 4 branches. Each branch contains a sequence of convolution layers with different kernel sizes and dilation rates, representing different receptive fields. We then integrate information from all branches to exploit comprehensive semantic information with a wider receptive field to determine the real foreground and background. Refer to the Supplemental for LSR implementation details.

**Feature-guided Loss**

Scribble-based methods often suffer from the lack of object information provided by the limited labeled data. Previous methods (Zhang et al. 2020b; Yu et al. 2021) exploit the information by using the pixel features of images, like colors and positions, assuming that foreground objects have visually distinctive features from backgrounds. However, in COD, such features are no longer a strong cue for boundary regions. It usually requires semantic information to decide the exact boundaries. Therefore, we design feature-guided loss based on both simple visual features (context affinity) and complex semantic features (semantic significance loss). As shown in Figure 5(b), semantic features extracted from the model respond actively to camouflaged boundaries and provide valuable guidance in these regions.

**Context Affinity Loss.** Nearby pixels with similar features tend to have the same class. Following previous methods
where $S(i), C(i)$ are the position $(x_i, y_i)$ and colors $(r_i, g_i, b_i)$ of pixel $i$. $\sigma_S, \sigma_C$ are hyperparameters. $D(i, j)$ calculates the probability of pixel $i, j$ having different classes ($P_{i,j}$ is the probability of positive labels for pixel $i, j$), and thus the context affinity loss $L_{ca}$ encourages visually dissimilar pixels to have different labels or vice versa:

$$D(i, j) = 1 - P_i P_j - (1 - P_i)(1 - P_j)$$

$$L_{ca} = \frac{1}{M} \sum_i \frac{1}{K_d(i)} \sum_{j \in K_d(i)} K_{vis}(i, j) D(i, j),$$

where $K_d(i)$ is a neighbor $n \times n$ regions ($n$ is set to 5 in our experiments) of center pixel $i$. Through context affinity loss, the model can quickly learn from the unlabeled pixels.

**Semantic Significance Loss.** In COD, pixels near boundaries usually resemble each other visually, and semantic features, especially those that distinguish segmented objects (thus significant), become crucial for the exact predictions. In this case, we design the semantic significance (SS) loss that utilizes significant features to refine the predictions of boundary regions.

Here, the SS loss is computed inside small boundary regions (in practice, we divide an image to 8 $\times$ 8 region blocks $(R_1, ..., R_r)$ with a step size of 4). A valid boundary region is defined as an area with at least 30% of the pixels being confidently classified as foreground or background (pixels with scribble annotation or model prediction above 0.8 is confidently classified). The design has two benefits. First, in non-boundary regions, low-level visual features suffice to provide good guidance. Second, it reduces the computation cost greatly. The semantic feature map $F_{ss}$ is extracted before the final prediction layer and its gradient is stopped (like the detach operation in Pytorch). The significance of a featured channel is determined by its covariance

$$L_{ss} = \frac{1}{M} \sum_i \frac{1}{K_d(i)} \sum_{j \in K_d(i)} K_{vis}(i, j) C(i) - C(j),$$

where $C(i)$ is the feature map of the input and its gradient is stopped (like the detach operation in Pytorch). The significance of a featured channel is determined by its covariance...
Figure 5: (a) shows that prediction on normal input is more accurate than on its transform. The design of the CV loss considers this reliability bias; (b) visualizes kernels of the visual features (VF) and semantic features (SS) in $K_{vis}$ and learnt semantic features (SS) in $K_{sem}$. Images are divided into 32×32 blocks (red blocks means boundary regions). We calculate the kernels with respect to the center pixels (anchors) inside blocks. White indicates high energy when the pixel label differs from the anchor.

with confidently classified predictions:

$$
    Sig_{i} = \text{cov}'(F_{ss,i}, P), i \in \{1, ..., C\},
$$

where $F_{ss,i}$ is the feature map of the $i$-th channel and $\text{cov}'$ means covariance, computed only on confidently classified pixels. The reason behind is that the above correlations roughly show how well the features distinguish the foreground and background. Low-significance features are unwanted since they may include the “camouflaged” parts of the object and confuse the model.

We then take the top $N$ channels ordered by $Sig$ to form significant feature map $F_{ss} \in \mathbb{R}^{H \times W \times N}$. In this task, we set $N$ to 16 to balance between performance and computation cost. The semantic significance loss has a similar formulation to context affinity loss:

$$
    K_{sem} = \exp\left(-\frac{||S(i) - S(j)||^2}{2\sigma_S^2} - \frac{||F_{ss}(i) - \hat{F}_{ss}(j)||^2}{2\sigma_C^2}\right),
$$

$$
    L_{ss} = w_{ss} \frac{1}{M} \sum_{k} \frac{1}{|R_k|} \sum_{i,j \in R_k} K_{sem}(i,j) D(i,j),
$$

where $S(i)$ is the position of the pixel, $R_k$ are valid boundary regions, and $w_{ss}$ is set to increase with the epoch number (exponential ramp-up to 0.15 in practice) since the model has not learned well-represented features at the beginning.

In conclusion, the feature loss $L_{ft}$ can be written as the sum of both loss in $L_{ft} = L_{ca} + L_{ss}$.

Consistency Loss

Weakly-supervised methods often suffer from inconsistent predictions. Similar to self-consistency mechanisms in self-supervision and weakly-supervision (Laine and Aila 2016; Mittal, Tatarchenko, and Brox 2019; Yu et al. 2021; Pan et al. 2021), we propose the cross-view (CV) consistency loss to alleviate the problem by minimizing the difference between the predictions of the input and its transform. Compared to others, the CV loss excels in that it considers the reliable difference. As shown in Figure 5(a), we observe that the model has more reliable output with normal input than transformed input, which is plausible considering more loss functions are computed on normal input. The proposed CV loss pushes the predictions to the reliable one and leads to a solid improvement in performance. In addition, the predictions tend to be uncertain due to visual similarity between background and foreground in COD, and we design an inside-view consistency loss to improve the stability of predictions.

Cross-View Consistency Loss. For a neural network function $f_{\theta}(\cdot)$ with parameter $\theta$, some transformations $T(\cdot)$, input $x$, the ideal situation is $f_{\theta}(T(x)) = T(f_{\theta}(x))$. Here, the transform includes combinations of resizing, flipping, translation and cropping, and is randomly chosen. The choice of it is explored in the ablation study. As regularization, we use the similar consistency loss $L_{cv'}$ (Yu et al. 2021) to push them towards each other:

$$
    Sm(p_1, p_2) = \frac{1 - \text{SSIM}(p_1, p_2)}{2},
$$

$$
    L_{cv'}(P_1, P_2) = \frac{1}{M} \sum_{i} (1 - \alpha) \cdot Sm(P_{1i}, P_{2i}) + \alpha |P_{1i} - P_{2i}|,
$$

where SSIM is single scale SSIM (Godard, Mac Aodha, and Brostow 2017). $p_1, p_2$ are two pixels. $\alpha$ is 0.85. $P_1, P_2$ are prediction maps of the input and its transform. $M$ is the total number of pixels and $i$ is a pixel index.

Considering the above-mentioned reliability bias, we aim for the predictions of the transform $\hat{P}$ to be pushed more than that of the normal input $P$. The key here is to weight their backward gradient differently, and the proposed cross-view consistency loss can be written as:

$$
    L_{cv} = (1 + \gamma) L_{cv'}(P_1, \hat{P}) + (1 - \gamma) L_{cv'}(P_1, \hat{P}_d),
$$

where $P_1, \hat{P}_d$ have the same values as $P, \hat{P}$ yet the gradient on them will be ignored during back-propagation (like the detach operation in PyTorch). If $\gamma = 0$, it is the original loss $L_{cv'}$; if $\gamma > 0$, the backward gradient that pushes $\hat{P}$ to $P$ is greater than the other way around, and thus the goal is reached. In practice, $\gamma$ is set to 0.3.

Inside-view Consistency Loss. We note that uncertain predictions are likely to be inconsistent. Therefore, we present the inside-view consistency (IV) loss which “looks” inside the output and encourages predictions with high certainty by minimizing their entropy. We also use a soft indicator to filter out noisy predictions: when the entropy is above a certain threshold, the prediction result is not sure and it is malicious to increase the certainty of the model in this case. The inside-view consistency loss is as below.

$$
    L_{iv} = w_{iv} \cdot \frac{1}{|I - B|} \sum_{(i) \in I - B} -P_{i} \log P_{i} - (1 - P_{i}) \log(1 - P_{i}),
$$
where $I, B$ are the set of all pixels and noisy pixels. $i$ is the pixel index, $w_{iy}$ is the weight of this loss and set to 0.05 in practice. The entropy threshold for the near-boundary pixel is set to 0.5 empirically. The loss is added in the late stage of training when predictions are relatively accurate.

Combined with all the consistency losses, we have the final consistency loss: $L_{\text{cons}} = L_{\text{cv}} + L_{ivy}$.

### Objective Function

Below is PCE loss, where $\hat{P}$ is the set of labeled pixels in the scribble map, $y_i$ is the true class of pixel $i$, and $\hat{y}_i$ are the predictions on pixel $i$: $L_{pce} = \frac{1}{N} \sum_{i \in \hat{P}} - y_i \log \hat{y}_i - (1 - y_i) \log (1 - \hat{y}_i)$. We compute all losses on main output $P$ while for the auxiliary outputs ($P_{1 \ldots 4}$), we compute only the PCE loss, inside-view consistency, and context affinity loss. $L_{\text{aux}} = L_{pce} + L_{iv} + L_{ivy} (i = 1, 2, 3, 4)$, where $L_{aux}$ is the loss function applied to the i-th auxiliary output. Here, we do not include the other two losses for their small improvement, possibly because they require high-level feature representations or accurate segmentation to guide the model. Every output is up-sampled by linear interpolation to the same size as the input. Finally, the total objective function of our output is: $L = L_{\text{cons}} + L_{f1} + L_{pce} + \sum_{i=1}^{4} \beta_i L_{aux}^i$, where $\beta_i = 1 - 0.2i$.

### Experiments

#### Datasets and Implementation Details

Our experiments are conducted on three SOD benchmarks, CAMO (Le et al. 2019), CHAMELEON (Skurowska et al. 2018), and COD10K (Fan et al. 2020a). Following previous studies, we relabel 4,040 images (3,040 from COD10K, 1,000 from CAMO) and propose the S-COD dataset for training. The remaining is for testing. We adopt four evaluation metrics: Mean Absolute Error (MAE), S-measure ($S_m$), and E-measure ($E_m$) ($F_\beta$) (Margolin, Zelnik-Manor, and Tal 2014). We implement our method with PyTorch and conduct experiments on a GeForce RTX2080Ti GPU. In the training phase, input images are resized to $320 \times 320$ with horizontal flips. We use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9, a weight decay of 5e-4, and triangle learning rate schedule with maximum learning rate 1e-3. The batch size is 16, and the training epoch is 150. It takes around 5 hours to train. As for the inference process, input images are only resized to $320 \times 320$. We then directly predict the final maps without any post-processing (e.g., CRF).

#### Comparison with State-of-the-arts

As we propose the first weakly-supervised method, we introduce 2 scribble-based weakly and 2 unsupervised SOD methods for comparison. We also provide the results of fully-supervised 8 COD and 12 SOD methods for reference. Quantitative comparisons are demonstrated in Table 1. Our method performs the best under four metrics on three benchmarks among weakly or unsupervised methods. It achieves an average enhancement of 11.0% on MAE, 3.2% on S-measure, 2.5% on E-measure, and 4.4% on weighted F-measure than the state-of-the-art method SCWSSOD (Yu et al. 2021). In addition, it outperforms 7 fully-supervised methods. We also find that
The second group, the ablation of consistency loss, shows improvements on all metrics except MAE. This indicates the benefit of the proposed consistency mechanism. The third group ablates our feature-guided loss. The final group is the overall component ablation of consistency loss and feature-guided loss. We see that both losses provide tremendous improvement in the test dataset.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Loss</th>
<th>MAE↓</th>
<th>$S_m$ ↑</th>
<th>$E_m$ ↑</th>
<th>$P_{m}^b$ ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/ pce</td>
<td>Baseline</td>
<td>0.215</td>
<td>0.612</td>
<td>0.633</td>
<td>0.387</td>
</tr>
<tr>
<td>w/ ft, iv</td>
<td>w/o cv</td>
<td>0.105</td>
<td>0.721</td>
<td>0.786</td>
<td>0.600</td>
</tr>
<tr>
<td></td>
<td>w/ cv(R)</td>
<td>0.097</td>
<td>0.727</td>
<td>0.807</td>
<td>0.629</td>
</tr>
<tr>
<td></td>
<td>w/ cv(R,F)</td>
<td>0.094</td>
<td>0.730</td>
<td>0.812</td>
<td>0.638</td>
</tr>
<tr>
<td></td>
<td>w/ cv(R,F,T)</td>
<td>0.094</td>
<td>0.730</td>
<td>0.808</td>
<td>0.637</td>
</tr>
<tr>
<td></td>
<td>w/ cv(R,F,T,C)</td>
<td>0.092</td>
<td>0.735</td>
<td>0.815</td>
<td>0.641</td>
</tr>
<tr>
<td>w/ ft</td>
<td>w/ cv'</td>
<td>0.095</td>
<td>0.723</td>
<td>0.801</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>w/ cv</td>
<td>0.095</td>
<td>0.726</td>
<td>0.804</td>
<td>0.632</td>
</tr>
<tr>
<td></td>
<td>w/ cs</td>
<td>0.092</td>
<td>0.735</td>
<td>0.815</td>
<td>0.641</td>
</tr>
<tr>
<td>w/ cs</td>
<td>w/ ca</td>
<td>0.095</td>
<td>0.727</td>
<td>0.807</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td>w/ ft</td>
<td>0.092</td>
<td>0.735</td>
<td>0.815</td>
<td>0.641</td>
</tr>
<tr>
<td>w/ pce</td>
<td>w/ cs</td>
<td>0.096</td>
<td>0.731</td>
<td>0.821</td>
<td>0.641</td>
</tr>
<tr>
<td></td>
<td>w/ ft</td>
<td>0.107</td>
<td>0.720</td>
<td>0.785</td>
<td>0.592</td>
</tr>
<tr>
<td></td>
<td>w/ cs, ft</td>
<td>0.092</td>
<td>0.735</td>
<td>0.815</td>
<td>0.641</td>
</tr>
</tbody>
</table>

Table 3: The ablation study for our loss functions on CAMO (Le et al. 2019). Groups correspond to ablations on transformations in cross-view consistency, on consistency loss, on feature loss, and on all losses. Here, pce stands for partial cross-entropy; ft and cs stand for feature-guided loss and consistency loss ($cs=cv+iv$, $ft=ca+ss$); cv and iv stand for cross-view and inside-view consistency loss; cv' means cross-view consistency without reliability bias; cv(·) specifies the transforms used in computing cv; R,F,T,C are resizing, flipping, translation and cropping.

Conclusion

In this paper, we propose the first weakly-supervised COD dataset with scribble annotation, which takes ~ 10 seconds to label an image (360 times faster than pixel-wise annotation). To overcome the weaknesses of current weakly-supervised learning and their application to COD, we propose a novel framework consisting of two loss functions and a novel network: a consistency loss, including consistency inside and cross images, regulates the model to have coherent predictions, and incline them to more reliable ones; a feature-guided loss locates the "hidden" foreground by comparing both manually computed visual features and learned semantic features of each pixel. The proposed network learns low-level contrast to expand scribbles to wider potential regions first and then analyzes logical semantic relation information to determine the real foreground and background. Experimental results show our method outperforms unsupervised and weakly-supervised state-of-the-arts with improvement, and is even competitive with the fully-supervised methods.
References


